Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Categorical variables and their effect

Year: The riders' count (dependent variable) has substantially increased year over year from 2018 to 2019

Season: Highest bike during summer and fall, likely due to favorable weather and temperatures and lower demand during spring and winter, potentially due to less favorable weather

Weather Situation: Riders' count is high during clear sunny or partial cloudy days. The count is moderately high during misty weather but falls considerably during rains

Month: The riders' count matches the seasonal pattern.

Holiday: Bike rentals are generally lower on holidays compared to regular working days, likely because fewer people commute.

Weekday: The median user count remains just about the same across the days in a week. The demand for bikes increases during weekdays and falls marginally over weekends.

Working day: The riders are marginally high on a working day, possibly due to the daily commutes.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

When all unique values of a variable are represented as dummies, one dummy can be perfectly predicted from the others (high multicollinearity). This creates ambiguity in feature selection and leads to an unstable or erroneous model.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

'Registered' has the highest collinearity or 0.95 with the target variable 'cnt', followed by 'casual' (0.67) and temp/atemp (0.63).

But considering the independent variables, atemp/temp has the highest correlation.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Validation is done using

1. Residual Analysis:

a. The predicted data is distributed randomly and does not follow any pattern

b. The residual distribution follows normal distribution with the mean of the residues is centered around 0 which ensures that there are no biases

This confirms a linear relationship between predictors and the target variable.

2. Evaluation on test data
a. The target values are predicted for the test data (y_data_pred) by using the model on the test variables (X_data_test).
b. Scatter plot of predicted target values vs actual values to ensure if the model explains the unseen data
c. Then the R^2 is calculated for the actual test data and the predicted data.
The R^2 score of test data matching or closer to the derived R^2 of the training data gives confidence in the model.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

1. atemp
2. Yr
3. LightRain

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>
1. Linear regression is a supervised learning algorithm used to predict a continuous target variable **y** which is affected by dependent or predictor variables (also known as features) $\mathbf{X_1, X_2, X_3 \ldots X_n}$
It is most suitable or applicable when the relationship between the dependent variable (y) and the predictor variables is linear. This can be figured out or validated by the residual analysis.

2. The relationship between the dependent and the predictor variables can be defined by a linear equation mentioned below

a. Simple Linear Regression equation (with only 1 predictor variable)
$y = \beta_0 + \beta_1 X_1 + \epsilon$

b. Multiple Linear Regression equation
$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots. + \beta_0 X_n + \epsilon$

Where,
β_0 = is the intercept on Y i.e the value when all the predictors are 0

$\beta_1, \beta_2, \ldots, \beta_n$ = are the weights/coefficients of each predictor variables which define the amount of change in y for 1 unit of change in a particular predictor variable X when the other predictors are 0

ϵ = error term or residual

2. The objective of the linear regression algorithm is to
a. Find out the predictors that are most significant in defining the target variable Y and
b. Determine weights/coefficients of their corresponding predictors such that the error is minimal
This is calculated by cost function
J = $\sum$(Residuals)^2, where *Residual = actual* y_i *– predicted* y_i

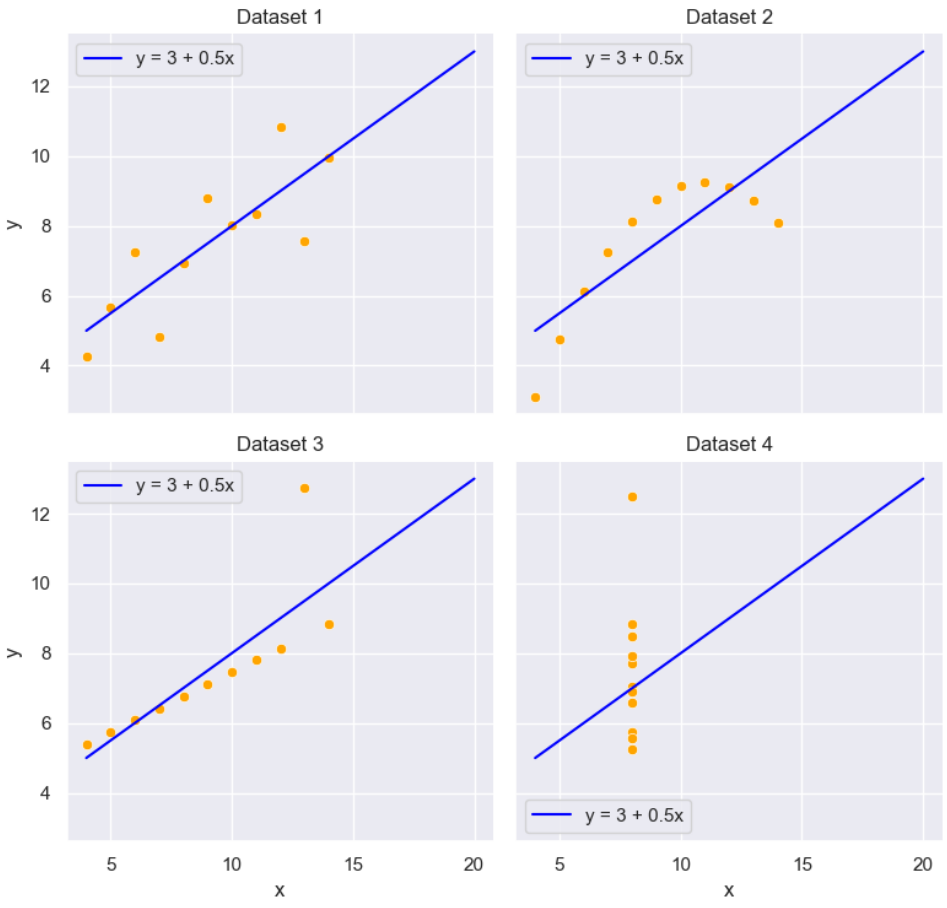
3. Linear regression can be applied or the relation between target and predictor variables is assumed to be linear when
a. The residuals of all the values follow a normal distribution about mean 0
b. The plot of predicted vs actual plot does not follow a random distribution with no visual pattern

---

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties (e.g., mean, variance, correlation, and regression line) but differ significantly in their visual patterns.

Given below is a sample plot of such data sets with identical stats and follow the model



Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's R measures the linear correlation between two variables. It is a statistical metric used to understand how strongly two variables are related and whether their relationship is positive or negative. It's value ranges from −1 to +1.

The positive sign indicates a positive correlation that is one variable increases in increase in another.
Negative sign indicates that increase in one variable results in decrease of another and vice versa
The magnitude of the value indicates the strength of the correlation. The higher the values the stronger the correlation

$$r = \left(\sum (x_i - \text{mean}(x))\,(y_i - \text{mean}(y)) \right) \Big/ \left(\sqrt{\left(\sum (x_i - \text{mean}(x))^2 \sum (y_i - \text{mean}(y))^2\right)} \right)$$

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling is the process of adjusting the values of features (variables) in the dataset so that they have comparable magnitudes. This ensures that no single feature dominates the model due to its large values.

For example, consider a model $y = 2X_1 + 500\,X_2$, which predicts the loan **y** a user can be granted for buying a house. It depends on feature $\mathbf{X_1}$, salary (ranging in thousands) and another $\mathbf{X_2}$, experience (say 1-30).

Here, the large values of salary can overpower the effect of experience in machine learning algorithms. Scaling resolves this issue.

Scaling is performed
a. To remove biases
b. To improve model performance and
c. For consistency

There are 2 types of scaling

1. **Normalize scaling** (MinMax scaling):
a. Used when the data is needed to be in a fixed range.
b. Scales all the numerical data between 0 and 1
c. Formula: scaled value $x' = (x - x_{min} / x_{max} - x_{min})$

2. **Standardised scaling:**

a. Used for scaling when data needs no range limit but needs to fit into a normal distribution
b. Scales data for features to be centred around mean 0 and have 1 standard deviation
b. Formula: scaled value x' = (x - μ) / σ

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

VIF stands for **Variance Inflation Factor**. It is used for identifying the degree of multicollinearity between the features used as predictors in a regression model.
It measures how much variance in the coefficients in a regression model are inflated due to the correlation between the predictor variables.

Formula: $VIF = 1 / (1-R^2)$,
where R^2 is the coefficient of determination which explains how of variance is the target variable is explained by the predictor

VIF becomes infinite when R^2 becomes 1, that is the variance of in target variable is 100% explained by the predictor variables. This is due to the perfect collinearity between the predictors.

Eg: In the given assignment, cnt = casual + registered. When casual riders are 0, cnt = registered and vice versa. Hence, change in one of them has significant variance in another. Therefore, if a model (hypothetically) model created using these predictors would change the coefficient and makes model unreliable.

When this happens, the coefficients of the predictors vary drastically when one of them changes, impacting the accuracy of the model.

Thus, VIF helps in identifying variables with high collinearity and removing them to make the model consistent. Variables with VIF > 5 are considered have high collinearity (80%)

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot (Quantile-Quantile plot) is a graph that helps us check if the data (e.g., residuals) follows a specific distribution, like a normal distribution. If the data matches the expected distribution, the points in the Q-Q plot will lie close to a straight diagonal line.

In linear regression, one assumption is that the residuals are normally distributed. A Q-Q plot is used to validate this assumption.

If the residuals are normally distributed, the points in the Q-Q plot will align with the diagonal line.

If the residuals are not normal, it can affect the reliability of your regression model.